SUPPORT VECTOR MACHINE FOR SOLVING IMBALANCED DATASET PROBLEM

ISMAIL BIN MOHD KHAIRUDDIN

A project report submitted in partial fulfilment of the
requirements for the award the degree of
Master of Engineering
(Electrical – Mechatronics and Automatic Control)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JANUARY 2012

To my beloved and supportive parent,
Hayati Binti Abu Bakar and Mohd Khairuddin Bin Shafie
All my siblings and my friends for their eternal support, encouragement and inspiration throughout my journey of education

ACKNOWLEDGEMENT

In the Name of Allah S.W.T, Most Gracious, Most Merciful. I am grateful to Allah for His guidance, and only by His strength I have successfully completed my master project and the write up on this thesis.

Firstly, I would like wish to express my gratitude and appreciation to my supervisor, Dr. Zuwairie Bin Ibrahim for his precious guidance, assistance, advice and positive comments throughout the accomplishment of this project.

I would like to express heartily and sincere thankfulness Mr Ibrahim Bin Shapiai who had taught and helps me a lot on Support vector Machine, and also to all my friends for the encouragement, cooperation and inspiration they gave all along the way to the completion of this project.

Last but not least, I would like to thank my beloved family for their determined support, encouragement and understanding. I am grateful to all these important peoples

ABSTRACT

Most of machine learning classifiers such as Neural Network (NN), Naïve Bayes, and Decision Tree Method C4.5 are failed to classify the data when it deals with imbalanced data set. This is because; most of classifiers are biased to the majority class, tend to ignore minority class and treat the minority class as a noise/disturbance/variance. Generally, to tackle the imbalanced data set problem it consists of two strategies which are data level and algorithm level. The data level method consist of features selection and re-sampling the data such as undersampling, oversampling and combination of both undersampling and oversampling, while for algorithm level it consist internal modification of learning program. In this project, the Support Vector Machine (SVM) classifier is proposed in order to investigate the imbalanced data set problem. The investigation is obtained by measured the performance based on SVM classifier. This investigation will cover and measure the performance SVM classifier by measuring the g-mean value. The performance of SVM classifier is measured by measuring the g-mean value .Therefore, in order to increase the performance of SVM classifier oversampling methods called SMOTE is introduced and combine with it and the g-mean value is calculated. Experimental validation on the proposed algorithm is performed and demonstrated on various set of imbalanced data sets. Some experiment have been design to validate the proposed algorithm and performed it with various set of imbalanced data sets. Finally, the result is for each proposed algorithm is being compared and analyze.

ABSTRAK

Hampir kesemua sistem mesin pengelas seperti *Neural Network* (NN), *Naïve Bayes*, serta *Decision Tree Method C4.5* gagal untuk mengelaskan data apabila berdepan dengan masalah ketidakseimbangan data. Hal ini kerana, pengelas memihak kepada kelas majoriti dan cenderung mengabaikan kelas minoriti malah menganggapnya sebagai gangguan. Secara amnya, untuk menyelesaikan masalah ketidakseimbangan data ia meliputi dua strategi iaitu tahap data dan tahap algoritma. Bagi, tahap data ianya meliputi pemilihan input dan re-sampling seperti undersamplin, oversampling dan juga kombinasi *undersampling* dan *oversampling*, sementara itu, tahap algoritma meliputi perubahan learning program. Dalam projek ini, pengelas *Support Vector Machine* (SVM) telah diperkenalkan untuk mengkaji masalah ketidakseimbangan data. Kajian ini diperolehi berdasarkan pengiraan prestasi SVM. Prestasi SVM di ukur dengan mengukur nilai *g-mean.* Selepas itu, teknik oversampling iaitu SMOTE di perkenalkan dengan harapan untuk meningkatkan bilangan kelas minoriti. Eksperimen untuk pengesahan terhadap algoritma yang di cadangkan telah dilakukan ke atas pelbagai ketidakseimbangan data. Akhir sekali, hasil daripada setiap algoritma yang di cadangkan, di banding dan dianalisis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	-	Support Vector Machine
MMC	-	Maximal Margin Classifier
SMOTE	-	Synthetic Minority Oversampling Technique
WE	-	Wilson's Editing Technique
ADASYN	-	An Adaptive Synthetic Technique
ANN	-	Artificial Neural Network
NB	-	Naïve Bayes
BVQ	-	Bayes Vector Quantizer
LVQ	-	Labeled Vector Quantizer
VQ	-	Vector Quantizer
AIS	-	Adaptive Inference System
OFS	-	Orthogonal Forward Selection
ROWLS	-	Regularized Orthogonal Weighted Least Square
GA	-	Genetic Algorithm
AUC	-	Area Under ROC Curse
FSVM	-	Fuzzy Support Vector Machine
SDC	-	SMOTE with Difference Cost
QP	-	Quadratic Programming
KKT	-	Karush-Kuhn-Tucker
G-mean	-	Geometric Mean

LIST OF SYMBOLS

S	-	Training sample
L	-	Training set size
N	-	Dimensional input space
$H_{optimal}$	-	Optimal hyperplane
ξ	-	Slack variable
w	-	Weight vector
b	-	Bias
α	-	Dual variable
L	-	Primal lagrangian
W	-	Dual lagrangian
C	-	Margin parameter
K	-	Nearest neighbor parameter

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction to Support Vector Machine (SVM) in Brief

Support Vector Machine (SVM) is a computer algorithm that learns by example to assign labels to object and have been developed by Vapnik and co-worker [3] and one of the most attractive development in classifier design. In general, there are two types of Support Vector Machine (SVM) which are Support Vector Machine Classification and Support Vector Machine Regression. In this thesis only Support Vector Machine Classification will be focused on. During classification process, the SVM classifier constructs the hyperplane or set of hyperplanes in a high or infinite dimensional space. The vector that lies near the hyperplane is called support vector.

Generally, there are three main concepts for SVM classifier called Maximal Margin Classifier (MMC), Soft Margin and also Kernel Approach. In SVM the MMC is the simplest and the main block for more complex SVM, while for Soft Margin, it allow some misclassification of training data plotting. For Kernel Approach, it normally used to change the dimensional of the training data.

SVM classifier has been applied in many applications such as face detection [24], image retrieval [23] and text categorization [25] and shown remarkable success in classification better than other classifiers.

1.2 Introduction to Imbalanced Datasets

In general the data set problem can be grouped into two types which called balanced data set and imbalanced data set. Balanced data set is occurs when the distribution in the dataset for each class are same. For the real world problems, the data set that occurs is not really balanced, called imbalanced data sets. Imbalanced data sets can be defined as data sets that the distributions of majority class which is larger than minority class.

In recent years, the class imbalance problem has been one of the emergence challenges in machine learning and can be found in various fields such as biomedical [4], remote-sensing [5], and engineering [6], computer-security [7], and manufacturing industries [8]. Generally, the imbalance can be grouped into the multi-class classification and binary class classification. For the binary class problem, the output consist two class which called positive class and negative class.

Generally, there are two types of imbalance data set can be found in binary classification. First is called as between-class imbalance and the other one is named within-class imbalance. For the type one, the imbalance data is defined as one class has much more example than the other class. Meanwhile, for the within-class some subset of the class has more example than the other subset of the same class [16].

1.3 Problem Statements

Recently, the imbalance data set problem has received a lot of interest of researchers in the Machine Learning community by the virtue of the fact that the performance of the learning algorithm degrades significantly when it deals with the imbalance data set. This is because; most of classifier that available is not able efficiently learns the imbalanced data set. The decision boundary establish by the classifier is tends to favor the majority class and tends to ignore the minority class and treat it as noise. Hence, to overcome this problem a learning algorithm called Support Vector Machine (SVM) will be introduced in order to deal the imbalanced data set problems

1.4 Objectives of Project

The goals of this project are:

- To investigate the imbalanced data set problem based on Support Vector Machine (SVM) classifiers. During the investigation, the performance measure is calculated.
- To investigate the imbalanced data set problem when Support Vector Machine (SVM) is combining with Synthetic Minority Oversampling Technique (SMOTE) algorithm. Then the performance measure is evaluated.

1.5 Scopes of Project

In the design of a Support Vector Machine (SVM) for solving imbalanced data set problem, the scope of the projects has been defined as follows:

1. The algorithm of SVM is computed into Matlab.
2. There are four types that will be considered as case study like Haberman's Survival Dataset, Pima Indian Diabetes, German Credit and Liver Disorder. This data is taken from UCI Machine Learning Repository [17].
3. Geometric mean (g-mean) will be used in order to measure the performance of this imbalance classier.

1.6 Outlines of the Thesis

Chapter 1 presents an overview of Support Vector Machine (SVM), the imbalanced data set, the objective of this project, the scope of the project and the problem statements. Chapter 2 gives an insight to the research and development classifier to solve the imbalance data set done by various researchers.

Chapter 3 will explain the methodology of how the investigation of imbalanced classifier is being proposed.

Chapter 4 mainly devoted for demonstrating the experimental results of the project, performance and discussion.

Chapter 5 deals with the summary and conclusions of the project. Some recommendation and suggestions for the future development of the project are also discussed.

1.7 Summary

This chapter is briefly introduced a SVM and imbalanced data set. Then the problem statements, objectives, scope of project and outlines of the thesis are presented and discussed. It is important to have those before attempting to start the project. The idea on the project has been briefly overviewed.

REFERENCES

1. Nello Cristianini and John Shawe-Taylor (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. UK: Cambridge University Press.
2. Tristan Fletcher (2009) Support Vector Machine Explained. *UCL*,
3. V. N Vapnik (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag
4. William S. Noble (2006). What is Support Vector Machine?. *Journal Nature Biotechnology*. Vol. 24 (No. 12), pp. 1565-1567
5. Giorgous Mountrakis, Jungho Im, and Caesar Ogole (2010). Support Vector Machine in Review, *ISPRS Journal of Photogrammetry and Remote Sensing*. Vol. 66, pp. 247-259
6. Giang H. Nguyen, Abdesselam Bouzerdoum, and Son L. Phung, (2008) A Supervised Learning Approach for Imbalanced Data Sets, *Proceeding 19th International Conference on Pattern Recognition.* pp. 1-4.
7. Wenjie Hu, Yihua Liao and V.Rao Vemuri (2003). Robust Support Vector Machines for Anomaly Detection in Computer Security. *Proceeding of International Conference on Machine Learning and Applications.*
8. WK Yip, KG Law, and WJ Lee (2007). Forecasting Final/Class Yield Based on Fabrication Process E-Test and Sort Data, *Proceeding of IEEE International Conference on Automation Science and Engineering*, pp. 477-483.
9. Rehan Akbani, Stephen Kwen, and Nathalie Japkowics (2004) Applying Support Vector Machines to Imbalanced Datasets. ECML. pp. 39-50.
10. Asrul Adam, Ibrahim Shapiai, Zuwairie Ibrahim, Marzuki Khalid, Lim Chun Chew, Lee Wen Jau, and Junzo Watada (2010). A Modified Artificial Neural

Network Learning Algorithm for Imbalanced Data Set Problem. *Second International Conference on Computational Intelligence*

11. Ting Yu, Tony Jan, Simeon Simoff and John Debenham (2007) A Hierarchical VQSVM for Imbalanced Data Sets. *Proceedings of International Joint Conference on Neural Networks*
12. Xiaohong Fan and Zongyao He (2010). A Fuzzy Support Vector Machine for Imbalanced Data Classification. *International Conference on Optoelectronics and Image Processing.*
13. Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla and Sven Krasser (2008). SVMs Modeling for Highly Imbalanced Classification. *IEEE Transactions on System, Man and Cybernetics*. Vol. 39, pp 281-288
14. Linde, Y. and Buzo, A. and Gray, R. M (1980). An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*. pp. 702-710.
15. Shuxue Zou, Yanxin Huang, Yan Wang, Jianxin Wang, Chunguang Zhou (2008) SVM Learning from Imbalanced Data by GA Sampling for Protein Domain Prediction. *The 9th International Conference for Young Computer Scientists*
16. Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, (2006) Handling imbalance datasets: A review, *GESTS International Transaction on Computer Science and Engineering*, Vol. 30.
17. Asuncion, A. & Newman, D.J (2007). UCI Machine Learning Repository. Retrieved on 2011, 10
18. Cheng G. Weng and Josiah Poon (2006). A New Evaluation Measure for Imbalance Datasets. *Proceedings of Seventh Australasian Data Mining Conference*
19. Kubat. M and Matwin, S (1997). Addressing the Curve of Imbalanced Training Set: One Sided Selection. *In Proceeding of the Fourteenth International Conference of Machine Learning*. pp. 179-186
20. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall. and W. Philip Kegelmeyer (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. Vol. 16, pp. 321-357.
21. Barandela, R., Sánchez, J.S., García, V., and Rangel, E (2003). Strategies for Learning in Class Imbalance Problems, *Pattern Recognition*, Vol. 36(No. 3), pp. 849-851.

22. Benjamin X. Wang and Natalie Japkowics (2009). Boosting Support Vector Machine for Imbalanced Dataset, *In Proceeding of Knowledge Inf System*, Vol. 25 (No. 1), pp. 1-20

23. Lei Zhang, Fuzong Lin and Bo Zhang (2001). Support Vector Machine Learning For Image Retrieval. *Proceeding of International Conference on Image Processing*

24. Yongmin Li, Shaogang Gong, Jamie Sherrah and Heather Liddell. Multi-view Face Detection Using Support Vector Machines and Eigenspace Modeling. *Proceeding of 4th International Conference on Kowledge-Based Intelligent Engineering Systems and Allied Technology*.

25. Joachims, T (1998). Text Categorization with SVM: Learning with Many Relevant Features. *Proceedings of ECML-98, 10th European Conference on Machine Learning.*

26. Xia Hong, Sheng Chen,and Chris J. Harris (2007). A Kernel-Based Two-Class Classifier for Imbalanced Data Sets. *IEEE Transactions on Neural Networks*. Vol. 18, (No. 1).

27. Alberto Fernández, María José del Jesus, Francisco Herrera (2009). On The Influence of an Adaptive Inference System in Fuzzy Rule Based Classification Systems for Imbalanced Data-Sets

28. Kaizhu Huang, Haiqin Yang, Irwin King, and Michael R. Lyu (2004) Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

29. Na Chu, Lizhuang Ma, Jing Li, Ping Liu and Yang Zhou (2010). Rough Set Based Feature Selection for Improved Differentiation of Traditional Chinese Medical Data, *Seventh International Conferece on Fuzzy Systems and Knowledge Discovery*, pp 2667-2772

30. Claudia Diamantini and Domenico Potena (2009) Bayes Vector Quantizer for Class-Imbalance Problem. *IEEE Transactions on Knowledge and Data Engineering*. Vol. 21 (No. 5)

31. Luiz Merschmann and Alexandre Plastino (2007). A Lazy Data Mining Approach for Protein Classification, *IEEE Transactons on Nanobioscience*, Vol. 6 (No. 1)

32. David Tian and Keith Burley (2010). Classification of Micro-Array Gene Expression Data Using Neural Networks. *International Joint Conference on Neural Networks*, pp. 1-8

33. S. Sivakumari, R. Praveena Priyadarsini, P. Amudha (2009). Performance Evaluation of SVM Kernels Using Hybrid PSO-SVM. *Journal ICGST-AIML*, Vol. 9 (1)